PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-251412
(43)Date of publication of application : 06.09.2002

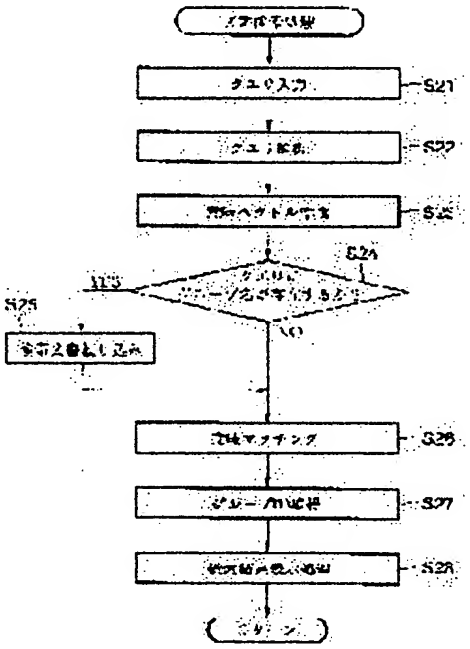
(51)Int.Cl. G06F 17/30
G06F 12/00

(21)Application number : 2001-047027
(22)Date of filing : 22.02.2001

(71)Applicant : CANON INC
(72)Inventor : FUNAKOSHI MASANOBU

(54) DOCUMENT RETRIEVING DEVICE, METHOD, AND STORAGE MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To reduce processing time in document retrieving processing, where multi-dimensional vector matching is used, even when a large amount of document data is retrieved.
SOLUTION: Document data retrieving is conducted using a document database in which pairs, each composed of document data and a semantic vector representing its contents with a multi-dimensional vector, are stored. Document data stored in the document database is classified into groups according to similar semantic vectors, and the groups are stored in a group table. When a query described in a natural language statement is entered, a language analysis is made for the query to generate the semantic vector of the query (S21-S23). Based on the language analysis result, a group is selected from the group table (S24, S25). For document data belonging to the selected group, matching between the semantic vector stored in the selected group and the semantic vector of the query is performed to retrieve for a document and then the result is output (S26-S28).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－251412
（P2002－251412A）
(43)公開日　平成14年９月６日(2002.9.6)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	17/30	３５０	Ｇ０６Ｆ 17/30	３５０Ｃ	５Ｂ０７５
		１７０		１７０Ａ	５Ｂ０８２
		３３０		３３０Ｃ	
	12/00	５２０	12/00	５２０Ｅ	

審査請求　未請求　請求項の数16　ＯＬ　（全 11 頁）

(21)出願番号　特願2001－47027(P2001－47027)

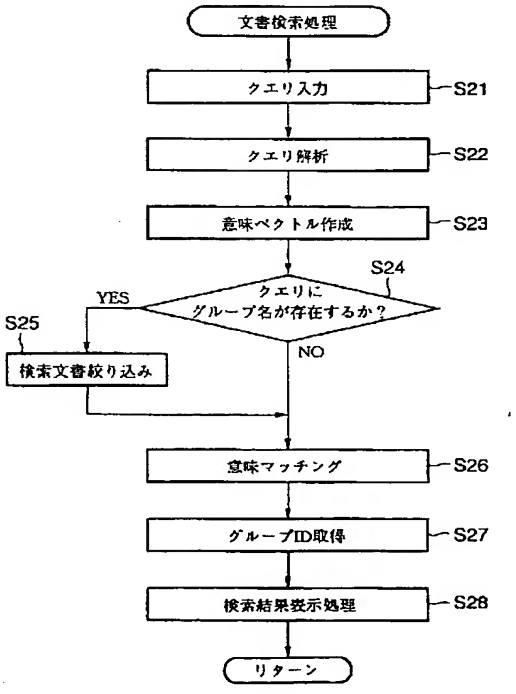
(22)出願日　平成13年２月22日(2001.2.22)

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号
(72)発明者　船越　正伸
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内
(74)代理人　100076428
弁理士　大塚　康徳　(外１名)

Ｆターム(参考)　5B075 ND03 ND04 NK06 NK32 NK43
PP02 PP14 PP24 PQ36 PQ46
PR06 QM05 UU06
5B082 AA00 EA07

(54)【発明の名称】　文書検索装置および方法ならびに記憶媒体

(57)【要約】
【課題】多次元ベクトルマッチングを用いた文書検索処理において、膨大な文書データが検索対象となった場合でもその処理時間を短く保つ。
【解決手段】文書データと、その内容を多次元ベクトルで表現する意味ベクトルとを対にして登録した文書データベースを用いて文書データの検索を行う。ここで、文書データベースに登録された文書データは、その意味ベクトルの近いグループに分類されてグループテーブルに登録される。自然言語文でクエリが入力されると、これを言語解析して、クエリの意味ベクトルを生成する(S21～S23)。そして、この言語解析の結果に基づいて、グループテーブルに登録されているグループを選択する(S24, S25)。そして、選択されたグループに属する文書データについて、データベースに登録された意味ベクトルとクエリの意味ベクトルとのマッチングを行って文書を検索し、結果を出力する(S26～S28)。



【特許請求の範囲】

【請求項１】　文書データと、その内容を多次元ベクトルで表現する意味ベクトルとを対にして格納する格納手段と、
前記格納手段に格納された文書データを、その意味ベクトルに基づいてグループに分類して管理する管理手段と、
検索条件として入力された自然言語文を言語解析して、検索条件文の意味ベクトルを生成する生成手段と、
前記言語解析の結果に基づいて、前記管理手段で管理されているグループを選択する選択手段と、
前記選択手段で選択されたグループに属する文書データについて、前記格納手段に格納された意味ベクトルと前記検索条件文の意味ベクトルとのマッチングを行って文書を検索する検索手段とを備えることを特徴とする文書検索装置。

【請求項２】　前記管理手段は、
前記格納手段によって格納された各文書データの意味ベクトルを用いて自己組織化マップを生成し、
前記格納手段に格納された各文書データの意味ベクトルに基づいて、前記自己組織化マップ上に各文書データを配置し、
前記自己組織化マップ上で近隣に配置された文書データで１つのグループを形成させることを特徴とする請求項１に記載の文書検索装置。

【請求項３】　前記管理手段は、各グループ毎に当該グループを代表する単語によるグループ名を割り当て、
前記選択手段は、前記言語解析処理の結果得られた、前記検索文が含む単語と一致するグループ名を有するグループを選択することを特徴とする請求項１または２に記載の文書検索装置。

【請求項４】　前記管理手段は、
グループに属する文書データの意味ベクトルに基づいて、当該グループの意味ベクトルを設定し、
前記グループの意味ベクトルに基づいて取得された単語を前記グループ名とすることを特徴とする請求項３に記載の文書検索装置。

【請求項５】　前記管理手段は、
前記グループの意味ベクトルに基づいて複数の単語を取得し、取得された複数の単語のうち、シソーラスツリー上で中間部に位置する単語を前記グループ名として選択することを特徴とする請求項４に記載の文書検索装置。

【請求項６】　前記検索手段による検索結果として、前記マッチングの度合が高い順に所定数の文書データを示す情報を出力する出力手段を更に備えることを特徴とする請求項１に記載の文書検索装置。

【請求項７】　前記出力手段は、前記文書データの各々に、当該文書データが属するグループを特定する情報を付加させて出力することを特徴とする請求項６に記載の文書検索装置。

【請求項８】　文書データと、その内容を多次元ベクトルで表現する意味ベクトルとを対にして格納する格納手段を用いた文書検索方法であって、
前記格納手段に格納された文書データを、その意味ベクトルに基づいてグループに分類して管理する管理工程と、
検索条件として入力された自然言語文を言語解析して、検索条件文の意味ベクトルを生成する生成工程と、
前記言語解析の結果に基づいて、前記管理工程で管理されているグループを選択する選択工程と、
前記選択工程で選択されたグループに属する文書データについて、前記格納手段に格納された意味ベクトルと前記検索条件文の意味ベクトルとのマッチングを行って文書を検索する検索工程とを備えることを特徴とする文書検索方法。

【請求項９】　前記管理工程は、
前記格納手段によって格納された各文書データの意味ベクトルを用いて自己組織化マップを生成し、
前記格納手段に格納された各文書データの意味ベクトルに基づいて、前記自己組織化マップ上に各文書データを配置し、
前記自己組織化マップ上で近隣に配置された文書データで１つのグループを形成させることを特徴とする請求項８に記載の文書検索方法。

【請求項１０】　前記管理工程は、各グループ毎に当該グループを代表する単語によるグループ名を割り当て、
前記選択工程は、前記言語解析処理の結果得られた、前記検索文が含む単語と一致するグループ名を有するグループを選択することを特徴とする請求項８または９に記載の文書検索方法。

【請求項１１】　前記管理工程は、
グループに属する文書データの意味ベクトルに基づいて、当該グループの意味ベクトルを設定し、
前記グループの意味ベクトルに基づいて取得された単語を前記グループ名とすることを特徴とする請求項１０に記載の文書検索方法。

【請求項１２】　前記管理工程は、
前記グループの意味ベクトルに基づいて複数の単語を取得し、取得された複数の単語のうち、シソーラスツリー上で中間部に位置する単語を前記グループ名として選択することを特徴とする請求項１１に記載の文書検索方法。

【請求項１３】　前記検索工程による検索結果として、前記マッチングの度合が高い順に所定数の文書データを示す情報を出力する出力工程を更に備えることを特徴とする請求項８に記載の文書検索方法。

【請求項１４】　前記出力工程は、前記文書データの各々に、当該文書データが属するグループを特定する情報を付加させて出力することを特徴とする請求項１３に記載の文書検索方法。

【請求項１５】　請求項８乃至１４のいずれかに記載の文書検索方法をコンピュータによって実現させるための制御プログラム。

【請求項１６】　請求項８乃至１４のいずれかに記載の文書検索方法をコンピュータによって実現させるための制御プログラムを格納する記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、言語解析による文書検索装置及び方法ならびにこれを記憶した媒体に関する。

【０００２】

【従来の技術】インターネットの爆発的な普及と共に、その入り口となるパーソナルコンピュータ（ＰＣ）や携帯情報機器（ＰＤＡ）の数も増大し、電子メールなどを利用した電子化文書によるコミュニケーションは我々の日常生活に浸透しつつある。このような状況において電子化文書の数は現在もなお増え続ける一方である。このため、膨大な文書データの中から必要な文書を迅速に検索する技術はますます重要視されており、より使いやすいインターフェースを持つ文書検索システムが登場している。

【０００３】この種の文書検索システムに、例えば、意味概念を用いた文書検索システムがある。これは、文書データベースにおいて、各文書データとその内容の意味概念を示す多次元ベクトルとを一対一に対応付けて格納しておき、ユーザが日常使っている自然言語を用いた簡単な質問文が入力されると、この質問文を言語解析してその意味概念を表す多次元ベクトルに変換し、文書データベース中の多次元ベクトルとマッチングをとることによって検索するものである。この手法によれば、ユーザが探している文書を自然言語を用いたインターフェースで検索することができる。

【０００４】

【発明が解決しようとする課題】しかしながら、多次元ベクトルのマッチングは、キーワードマッチングなどと比較すると計算量が大きくなるため、検索対象となる文書データの数が膨大になると、検索処理に多大な時間がかかってしまうという課題があった。

【０００５】また、多次元ベクトルのマッチングによる検索では、場合によっては検索結果の中に全く関係のない文書が混じってしまい、検索結果の精度が上がらないという課題もあった。

【０００６】本発明は、上記の課題に鑑みてなされたものであり、多次元ベクトルマッチングを用いた文書検索処理において、膨大な文書データが検索対象となった場合でもその処理時間を短く保つことを目的とする。

【０００７】

【課題を解決するための手段】上記の課題を達成するための本発明による文書検索装置は以下の構成を備える。すなわち、文書データと、その内容を多次元ベクトルで表現する意味ベクトルとを対にして格納する格納手段と、前記格納手段に格納された文書データを、その意味ベクトルに基づいてグループに分類して管理する管理手段と、検索条件として入力された自然言語文を言語解析して、検索条件文の意味ベクトルを生成する生成手段と、前記言語解析の結果に基づいて、前記管理手段で管理されているグループを選択する選択手段と、前記選択手段で選択されたグループに属する文書データについて、前記格納手段に格納された意味ベクトルと前記検索条件文の意味ベクトルとのマッチングを行って文書を検索する検索手段とを備える。

【０００８】また、上記の課題を達成するための本発明による文書検索方法は、文書データと、その内容を多次元ベクトルで表現する意味ベクトルとを対にして格納する格納手段を用いた文書検索方法であって、前記格納手段に格納された文書データを、その意味ベクトルに基づいてグループに分類して管理する管理工程と、検索条件として入力された自然言語文を言語解析して、検索条件文の意味ベクトルを生成する生成工程と、前記言語解析の結果に基づいて、前記管理工程で管理されているグループを選択する選択工程と、前記選択工程で選択されたグループに属する文書データについて、前記格納手段に格納された意味ベクトルと前記検索条件文の意味ベクトルとのマッチングを行って文書を検索する検索工程とを備える。

【０００９】

【発明の実施の形態】以下、添付の図面を参照しながら本発明の好適な実施形態を説明する。

【００１０】本実施形態では、文書データを登録する際に、文書の意味内容を示す多次元ベクトルを作成し、この多次元ベクトルを利用して予め文書を意味概念によって自動分類して管理する。そして、検索時のユーザクエリにこの分類と合致する言葉が現れた場合に、まず検索対象をこの分類に絞り込み、多次元ベクトルマッチングを行う。この検索対象の絞り込みにより検索結果の精度を高めると同時に、文書検索時の計算量を減らし、正確で迅速な文書検索を実行する。以下、本実施形態について詳細に説明する。

【００１１】図１は本実施形態による文書検索処理を実行するコンピュータ装置１００の構成を示すブロック図である。図１の構成において、ＣＰＵ１０１はマイクロプロセッサであり、文書検索処理のための演算、論理判断等を行い、ＰＣＩバス１０２を介して接続された各構成要素を制御する。ＰＣＩバス１０２はＣＰＵ１０１の制御の対象とする構成要素を指示するアドレス信号を転送し、ＣＰＵ１０１の制御の対象とする各構成要素のコントロール信号を転送し、各構成機器相互間のデータ転送を行う。

【００１２】ＲＯＭ１０３は読出し専用の固定メモリで

ある。ＲＯＭ１０３には本実施形態の構成における基本Ｉ／Ｏプログラムが格納される。また、ＲＡＭ１０４は、書込み可能のランダムアクセスメモリであって、各構成要素からの各種データの一時記憶と、本実施形態における各種処理が記述されたプログラムが格納され、このプログラムに基づいてＣＰＵ１０１が各種処理を行う。

【００１３】ＤＶＤＤ１０５はＤＶＤドライブである。ＤＶＤメディア（ＤＶＤ－ＭＥＤＩＡ）１０６に記録されているプログラムやデータはこのＤＶＤドライブ１０５を通じて本システムにロードされる。また、ＤＩＳＫ１０７に蓄えられた各種データをＤＶＤドライブ１０５を通じてＤＶＤ－ＭＥＤＩＡ１０６に書き込むことができる。なお、ＤＶＤ－ＭＥＤＩＡ１０６は、具体的にはＤＶＤ－ＲＯＭ，ＤＶＤ－ＲＡＭ，ＤＶＤ－Ｒ，ＤＶＤ－ＲＷ，ＤＶＤ－ＶＩＤＥＯ，ＤＶＤ－ＡＵＤＩＯなどのＤＶＤ規格のメディアを総称したものである。本実施形態において、ＤＶＤ－ＭＥＤＩＡ１０６は文書データやその関連データ、もしくはプログラムなどの大容量データの読み書きに用いられる。

【００１４】ＩＮＰＵＴＣ１０８は入力コントローラである。キーボード（ＫＢ）１０９やポインティングデバイス（ＰＤ）１１０から送られてくる入力信号は、このコントローラによって適宜適切な信号に変換された後、ＰＣＩバス１０２を経由してＣＰＵ１０１に送信される。

【００１５】ＫＢ１０９は、アルファベットキー、ひらがなキー、カタカナキー等の文字記号入力キー、及び、カーソル移動を指示するカーソル移動キー等のような各種のファンクションキーを備えている。ＰＤ１１０は、マウスやトラックボールなどのポインティングデバイスであり、表示画面上のカーソルやボタンなどを指摘するために使用される。

【００１６】ＤＩＳＫ１０７はデータやプログラム等を記憶するための外部メモリである。データやプログラム等は必要に応じて保管され、また、保管されたデータやプログラムはキーボードの指示により、必要な時に呼び出される。本実施形態における文書データベースは主にこのＤＩＫＳ１０７上に実装される。

【００１７】ＶＩＤＥＯ１１１はビデオコントローラである。ＰＣＩバス１０２を介して表示用のデータがここに蓄えられるとともに、表示用の信号に変換されて表示装置（ＤＩＳＰ）１１２に出力される。ＤＩＳＰ１１２には、陰極線管や液晶などが用いられ、各種処理の結果や装置の状態、ユーザに対するメッセージなどを表示する。

【００１８】ＤＥＶＣ１１３はデバイスコントローラである。ＰＣＩバス１０２を介して伝達されるＣＰＵ１０１の指示によって、このコントローラに接続されている機器を制御し、また、接続されている機器が出力する信号やデータをＰＣＩバス１０２を介してＣＰＵ１０１やＤＩＳＫ１０７に適宜伝達する。ＳＣＡＮ１１４はスキャナーである。ＤＥＶＣ１１３からの指示によって、光学的な方法によってここにセットされた原稿をスキャンし、原稿画像を読み取り、これをＤＥＶＣ１１３に出力する。

【００１９】ＮＩ１１５はネットワークインターフェースであり、本実施形態の文書検索システムをＬＡＮやインターネット１１６などを経由して外部のシステムと接続するための機器である。本実施形態の文書検索システムは、この接続を経由して、信号やデータを外部のシステムと送受信することが可能である。

【００２０】ＭＩＸ１１７はミキサーである。ＰＣＩバス１０２を介して音声出力用のデータがここに送られると、ＭＩＸ１１７はこれらの信号を合成しかつ音声出力用の信号に変換してスピーカ（ＳＰＫ）１１８に出力する。ＳＰＫ１１８は、処理結果や装置の状態、ユーザに対するメッセージ、音楽などを音声で出力する。

【００２１】かかる各構成要素からなる本実施形態のコンピュータ装置１００においては、キーボード１０９やポインティングデバイス１１０からの各種の入力に応じて各種処理を実行させる。すなわち、キーボード１０９やポインティングデバイス１１０から入力信号が供給されると、ＩＮＰＵＴＣ１０８を経由して、インタラプト信号がＣＰＵ１０１に送られ、ＣＰＵ１０１がＲＯＭ１０３内に記憶してある各種の制御信号を読み出し、それらの制御信号に従って、各種の制御が行われる。

【００２２】本実施形態では、コンピュータが基本Ｉ／Ｏプログラム、ＯＳ、及び本文書検索処理プログラムをＣＰＵ１０１が実行することによって、文書検索装置として動作する。基本Ｉ／ＯプログラムはＲＯＭ１０３中に書き込まれており、ＯＳはＤＩＳＫ１０７に書き込まれている。そして、本システムの電源がＯＮにされると、基本Ｉ／Ｏプログラム中のＩＰＬ（イニシャルプログラムローディング）機能によりＤＩＳＫ１０７からＯＳがＲＡＭ１０４に読み込まれ、ＣＰＵ１０１によるＯＳの動作が開始される。なお、文書検索処理プログラムは、図９～図１５のフローチャートによってに示される文書検索処理手順をＣＰＵ１０１によって実現させるためのプログラムコードである。

【００２３】図２は、本文書検索処理プログラム及び関連データをＤＶＤ－ＭＥＤＩＡ１０６に記録したときのＤＶＤ－ＭＥＤＩＡ１０６の内容の構成図である。本実施形態において、文書検索処理プログラム及び関連データはＤＶＤ－ＭＥＤＩＡ１０６に記録されている。図示のようにＤＶＤ－ＭＥＤＩＡ１０６の先頭領域には、このＤＶＤ－ＭＥＤＩＡのボリューム情報２０１とディレクトリ情報２０２が記録されており、その後にこのＤＶＤ－ＭＥＤＩＡ１０６のコンテンツである本実施形態の文書検索処理プログラム（実行ファイル）２０３と、文

書検索処理関連データ２０４が記録されている。

【００２４】図３はコンピュータ装置１００と本文書検索処理プログラムが記録されたＤＶＤ－ＭＥＤＩＡ１０６の模式図である。ＤＶＤ－ＭＥＩＤＡ１０６に記録された文書検索処理プログラム２０３および関連データ２０４は、図３に示したようにＤＶＤ－ＭＥＤＩＡドライブ（ＤＶＤＤ）１０５を通じて本システムにロードすることができる。すなわち、このＤＶＤ－ＭＥＤＩＡ１０６をＤＶＤＤ１０５にセットすると、ＯＳ及び基本Ｉ／Ｏプログラムの制御のもとに本文書検索処理プログラムおよび関連データがＤＶＤ－ＭＥＤＩＡ１０６から読み出され、ＲＡＭ１０４にロードされて動作可能となる。

【００２５】図４は、本文書検索処理プログラムがＲＡＭ１０４にロードされ実行可能となった状態のメモリマップを示す。コンピュータ装置１００の立ち上げ時にＲＯＭ１０３やＤＩＳＫ１０７よりロードされる基本Ｉ／Ｏプログラム４１１とＯＳ４１２が格納されている。また、ＤＶＤ－ＭＥＤＩＡ１０６からロードされた文書検索処理プログラム２０３とその関連データ２０４がそれぞれ文書検索処理プログラム４１３、関連データ４１４として格納される。また、ワークエリア４１５には、意味概念辞書ＭＤＩＣ４２１、検索結果バッファＲＢＵＦ４２２、結果出力数ＣＯＵＮＴ４２４、グループテーブルＧＴＢＬ４２３が存在している。

【００２６】図５は、本実施形態における意味概念辞書ＭＤＩＣの構成例を説明した図である。

【００２７】本実施形態における意味概念辞書（ＭＤＩＣ）４２１には、図示したように、単語ＩＤ５０１と、単語表記５０２と、その意味表現である多次元ベクトル５０３のリストで構成される。図５において、表の行が単語一個分のデータに相当する。このうち、単語ＩＤ５０１は、本実施形態の文書検索システムにおいて、各単語を識別管理するために、各単語に対して一意に割り振られている番号である。また、単語表記５０２は、各単語の表記を表す文字列である。また、意味ベクトル５０３は、各単語が持つ意味概念を多次元ベクトルで表現したものであり、予め定められている。

【００２８】なお、本実施形態における意味ベクトルの各次元は、名詞のシソーラスなどを参照して上位概念に相当する単語を適宜抽出して設定する。また、各単語の意味ベクトルは国語辞書などの語義文を言語解析して、各次元として選択された概念の合成に落とし込むことで作成するが、このような手法は意味処理として一般的であり、公知であるので、ここでは詳述しない。

【００２９】意味概念辞書（ＭＤＩＣ）４２１は以上の構成によって、単語ＩＤ５０１もしくは単語表記５０２をキーにして検索され、格納されている各情報を参照することが可能である。本実施形態において、意味概念辞書（ＭＤＩＣ）４２１は、後述する各処理において適宜検索、参照される。

【００３０】図６は、本実施形態における検索結果バッファ（ＲＢＵＦ）の構成例を説明した図である。本実施形態における検索結果バッファ（ＲＢＵＦ）４２２には、文書検索処理において検索されたデータのうち、意味マッチングの度合いが高いデータから順に結果出力数ＣＯＵＮＴ４２４の数だけ格納される。図５は、ＣＯＵＮＴ＝１０の場合の検索結果バッファ（ＲＢＵＦ）４２２の構成例である。

【００３１】図６に示したように、一つの検索結果データは、順位６０１と、意味近似度６０２と、意味ベクトル６０３と、文書ＩＤ６０４と、グループＩＤ６０５によって構成される。このうち、順位６０１は、ＲＢＵＦ４２２に格納されたデータの中における意味近似度が大きい順番である。また、意味近似度６０２は、各文書データに対応付けられた意味ベクトル６０３と、クエリを言語解析して作成した意味ベクトルとのマッチングを取ったときの度合いであり、本実施形態においてはパーセントで格納される。なお、２つの意味ベクトルのマッチングの度合いは、主にこれら２つのベクトルのなす角を算出することによって決定される。

【００３２】意味ベクトル６０３は、各文書データ中のテキストを言語解析して意味情報を抽出し、多次元ベクトルで表現したものであり、登録された各文書について予めデータベース中に格納されている。この意味ベクトル６０３は、通常、テキストを形態素解析を用いて品詞分解し、普通名詞や固有名詞を抜き出し、これらの持つ意味ベクトルを意味概念辞書（ＭＤＩＣ）５０２から検索して取得後、構文情報によって重み付けして合成することによって作成される。なお、このような手法は自然言語処理分野において公知であるので、ここでは詳述しない。

【００３３】また、文書ＩＤ６０４は、本実施形態の文書データベースに格納されている文書のＩＤであり、このＩＤを利用して実際の文書データを読み出すことが可能である。また、グループＩＤ６０５は、各文書が属する意味近似グループのＩＤであり、これを用いてグループテーブル（ＧＴＢＬ）４２３を参照することにより、この文書のグループ情報を取得することができる。なお、意味近似グループについては図７を参照して以下に説明する。

【００３４】図７は、本実施形態におけるグループテーブル（ＧＴＢＬ）の構成例を説明した図である。図７に示したように、本実施形態におけるグループテーブル（ＧＴＢＬ）４２３は、グループＩＤ７０１と、グループ名７０２と、データ数７０３と、文書ＩＤリスト７０４と、グループの意味ベクトル７０５によって構成される。このうち、グループＩＤ７０１は、本実施形態における文書データベースに登録してある文書データを、後述する手法によって意味的に分類することによって作成されるグループに一意に割り振られる番号である。な

お、このグループＩＤによって特定される１つのグループが１つの意味近似グループを構成する。このグループＩＤ７０１はＲＢＵＦ４２２のグループＩＤ６０５に対応している。また、グループ名７０２は、本実施形態における各意味近似文書グループに与えられたユニークな名前である。このグループ名を決定する処理は図１５を用いて後述する。また、データ数７０３は、このグループに属している文書データの数である。

【００３５】また、文書ＩＤリスト７０４は、このグループに属する文書データのＩＤのリストである。グループの意味ベクトル７０５は、当該グループに属する文書データの全ての意味ベクトルとのなす角度が最小となる多次元ベクトルである。グループテーブル（ＧＴＢＬ）４２３は、後述する意味近似グループ作成処理によって作成され、主に文書検索時の検索対象データの絞り込みに使用される。図７は、ＰＣ関連のニュース記事を自動分類した場合のグループテーブルＧＴＢＬの一例を示している。

【００３６】図８は、本実施形態における文書データベースに格納されている一つのデータ構造の構成例を説明した図である。図８に示したように、本実施形態における文書データベース中の１つのデータは、文書ＩＤ８０１、文書データポインタ８０２、キーワード８０３、意味ベクトル８０４によって構成される。

【００３７】ここで、文書ＩＤ８０１はこのデータ構造そのもののインデックス番号であると同時に、データベース中の文書データ自体のインデックス番号である。文書ＩＤ８０１はＲＢＵＦ４２２における文書ＩＤ６０４、ＧＴＢＬ７０２における文書ＩＤリストで用いられる文書ＩＤである。また、文書データポインタ８０２は、実際の文書データを指すポインタであり、このポインタを利用することにより実際の文書データにアクセスすることができる。また、キーワード８０３は、文書データ中に現れる代表的な単語のリストである。また、意味ベクトル８０４は文書データの意味を多次元ベクトルで表現したもので、ＲＢＵＦ４２２に格納される意味ベクトル６０３と同じものである。このようなデータ構造を取ることによって、実際の文書データにアクセスすることなく検索を高速に行うことが可能となる。

【００３８】以上の構成を備えた本実施形態の文書検索装置の動作について、以下、詳細に説明する。

【００３９】図９は、本実施形態における文書検索処理の全体を説明するフローチャートである。ステップＳ１は、ユーザのＫＢ１０９もしくはＰＤ１１０の操作を受け付け、これをコマンドとして解釈するユーザコマンド入力処理である。このような処理はＰＣなどを利用したシステムにおいて一般的なユーザインターフェース処理であり、公知であるのでここでは詳述しない。処理を終えると、ステップＳ２へ進む。

【００４０】ステップＳ２は、ステップＳ１で解釈したコマンドを判定して、各種処理に分岐するユーザコマンド判定処理である。ユーザコマンドが文書登録を指示している場合は、ステップＳ３へ進む。また、ユーザコマンドが文書検索を指示している場合は、ステップＳ４へ進む。また、ユーザコマンドがその他の処理を指示している場合は、ステップＳ５へ進む。

【００４１】ステップＳ３では、検索対象となる文書データベースに新たに文書データを登録する文書登録処理を行う。この処理は、図１０のフローチャートを用いて後述する。文書登録処理を終えると、ステップＳ１へ戻る。ステップＳ４では、ユーザクエリを受け付けて、そのクエリに合致するデータベース中の文書データを検索する文書検索処理を実行する。この文書検索処理は、図１１のフローチャートを用いて後述する。処理を終えると、ステップＳ１へ戻る。ステップＳ５では、表示のカスタマイズ、システムの各種設定変更、ユーザ情報登録などの、その他の処理を行う。これらの処理はこのようなシステムにおいて一般的であり、公知であるので詳述しない。処理を終えると、ステップＳ１へ戻る。

【００４２】なお、本実施形態のシステムにおいて、ステップＳ１へ処理が進むとき、新規登録文書数がある一定値を超えた、グループ情報の更新指示がされたなどの、諸処の条件が満たされた場合、システムに割り込みがかかり、ステップＳ６へ処理が進む。ステップＳ６は、その時点で本実施形態の文書データベースに登録されている文書をその意味によって新たにグループ分けして、図７に示したようなグループテーブルＧＴＢＬを作成する、意味近似文書グループ作成処理を行う。なお、この処理の詳細は、図１２のフローチャートを用いて後述する。意味近似文書グループ作成処理を終えると、ステップＳ１へ処理が戻り、割込みから復帰する。

【００４３】図１０は、図９のステップＳ３における文書登録処理を詳細化したフローチャートである。本処理により、図８に示す形態で文書が文書データベースに登録されることになる。

【００４４】ステップＳ１１では、ユーザが登録する文書を指定する登録文書指定処理を実行する。この処理は、実際にはＧＵＩを利用した文書ファイルのドラッグ・ドロップであったり、あるいは、登録ファイル名のリストを格納したテキストファイルの指定でも良い。この種の処理は本実施形態のようなシステムのユーザインターフェースとして極めて一般的であるので、詳述しない。処理を終えると、ステップＳ１２へ進む。

【００４５】ステップＳ１２では、ステップＳ１１で指定された文書データのうち、テキストの主要部分を抽出して言語解析する主要文解析処理が行われる。ここで、テキスト主要部分とは、テキストの大意が記述されている可能性が高い部分を抽出したものであり、文書の論理、レイアウト構造を利用して経験的に抽出される。主要文解析処理を終えると、ステップＳ１３へ進む。

【００４６】ステップＳ１３では、ステップＳ１２で得られた言語解析（主要文解析処理）の結果を利用して、この文書において特徴的に現れるキーワードを、重み付けして抽出するキーワード抽出処理が行われる。このようなキーワード抽出処理は言語処理の分野で一般的に行われており、公知であるので詳述しない。キーワード抽出処理を終えると、ステップＳ１４へ進む。

【００４７】ステップＳ１４では、ステップＳ１３で得られたキーワード情報を元に、この文書データの意味概念を表す多次元ベクトルを作成する文書意味ベクトル作成処理が行われる。本実施形態において、文書意味ベクトルの作成は、以下のように行われる。まず、ステップＳ１３で得られた各キーワードを意味概念辞書ＭＤＩＣ４２１（図５）から検索し、各キーワードの意味ベクトルを取得する。次に、各意味ベクトルにステップＳ１３で得られた重み付けを施した後、全ての意味ベクトルを合成する。処理を終えると、ステップＳ１５へ進む。

【００４８】ステップＳ１５では、ステップＳ１３で得られたキーワード、ステップＳ１４で得られた意味ベクトルの情報から図８に示した文書データを作成して文書データベースに登録する。なお、文書ＩＤは、重複が発生しないように、当該システムによって自動的に割り振られる。また、同時に、実際の文書データ自体もデータベース上の別領域に登録され、このときに文書データポインタが得られ、データベースに登録される。データ登録処理を終えると、ステップＳ１６へ進む。

【００４９】ステップＳ１６では、ステップＳ１４で作成した意味ベクトルを利用して、登録する文書が属するべき意味近似文書グループを決定する。意味近似文書グループの決定は、登録する文書の有する意味ベクトルと、グループテーブル（ＧＴＢＬ）中の各グループの意味ベクトル７０５とのマッチングを取り、最も近似しているグループを決定することによって行われる。文書の属するグループが決定されると、図７で示したグループテーブル（ＧＴＢＬ）中の決定されたグループの文書ＩＤリストに、ステップＳ１５で割り振られた文書ＩＤが格納される。処理を終えると、ステップＳ１７へ進む。

【００５０】ステップＳ１７は、これまでの登録処理の結果をユーザに通知する登録結果通知処理である。例えば、これは結果を記載したダイアログボックスの表示であったり警告音の出力であるが、この種の処理はユーザインターフェースとして一般的であり、公知であるので詳述しない。処理を終えると、文書登録処理を終了する。

【００５１】図１１は、図９に示したステップＳ４の文書検索処理の詳細を示すフローチャートである。

【００５２】ステップＳ２１では、文書を検索するためのクエリ（質問文）をユーザに入力してもらうクエリ入力処理を実行する。クエリは、ユーザが探している文書の内容を表現する簡単な文であり、例えば、「ＰＣの価格下落について」のように自然言語にて入力する。このような処理は一般の検索処理において公知であり、ここでは詳細な説明は行わない。クエリ入力処理を終えると、ステップＳ２２に進む。

【００５３】ステップＳ２２では、ステップＳ２１で入力されたクエリを言語解析するクエリ解析処理が実行される。この処理では、ステップＳ２１で入力されたクエリに、形態素解析、構文解析等の言語処理を行い、後の処理で利用しやすい形式である言語情報に変換する。なお、このような処理は一般の言語処理において公知であり、ここでは詳細な説明は行わない。クエリ解析処理を終えると、ステップＳ２３に進む。

【００５４】ステップＳ２３では、ステップＳ２２の結果得られた言語情報を利用して、クエリの意味ベクトルを作成する意味ベクトル作成処理が行われる。この意味ベクトル作成処理では、文書の意味ベクトルを作成する場合（ステップＳ１４）と同様に、クエリ中に現れる単語の意味ベクトルを意味概念辞書（ＭＤＩＣ）４２１から検索して取得し、これらを構文情報によって重み付けしてから合成することで作成する。意味ベクトル作成処理を終えると、ステップＳ２４へ進む。

【００５５】ステップＳ２４は、ステップＳ２２で得られた言語情報を利用して、クエリ中にグループテーブル（ＧＴＢＬ）４２３に格納されている意味近似グループのグループ名が存在しているかどうか調べる処理である。この処理の結果、クエリ中に１つ以上のグループ名が存在する場合はステップＳ２５へ進む。グループ名が１つも存在しない場合はステップＳ２６へ進む。

【００５６】ステップＳ２５では、ステップＳ２４の結果、クエリ中に存在するグループ名を持つ意味近似グループに属する文書のみに検索対象となる文書を絞り込む。検索対象をこの時点で絞り込むことによって、後の意味マッチング処理における処理量（計算量）を減らし、検索を高速化することができる。検索文書の絞込み処理が終わるとステップＳ２６へ進む。

【００５７】ステップＳ２６では、ステップＳ２３において得られたクエリの意味ベクトルと、文書データベース中の意味ベクトルとのマッチングを行って、ユーザが探している文書を検索する意味マッチング処理が実行される。意味ベクトルのマッチングは、前述したように２つの多次元ベクトルのなす角を算出することによって行われる。なお、ステップＳ２５によって検索対象となる文書が絞り込まれた場合は、検索対象となっている文書の意味ベクトルとのみマッチングが行われる。この処理の結果として、マッチング結果バッファ（ＲＢＵＦ）４２２には、データベース中のデータの中で、マッチングの度合いが上位のものから順に、結果出力数ＣＯＵＮＴの数だけ検索結果データが格納される。なお、このようは処理は一般のデータベース検索処理において公知であり、ここでは詳細な説明は行わない。以上の意味マッチ

ング処理を終えると、処理はステップＳ２７に進む。

【００５８】ステップＳ２７では、ステップＳ２６によって出力された検索結果バッファ（ＲＢＵＦ）中の各文書データが属するグループのＩＤを、グループテーブル（ＧＴＢＬ）４２３を検索することによって取得する、グループＩＤ取得処理を実行する。この処理によって取得したグループＩＤは、検索結果バッファ（ＲＢＵＦ）４２２中のグループＩＤ６０５として格納される。処理を終えると、ステップＳ２８へ進む。

【００５９】そして、ステップＳ２８では、ステップＳ２７によって作成された検索結果バッファＲＢＵＦ中の検索結果データに基づいて、ディスプレイに文書データを出力する検索結果表示処理である。この種の処理は表示を行うシステムにおいて一般的に行われており、公知であるので詳述しない。処理を終えると、文書検索処理を終了する。

【００６０】図１２は、図９のステップＳ６に示した意味近似文書グループ作成処理の詳細を示すフローチャートである。この処理は、各文書の多次元ベクトルを利用して予め文書を意味概念によって自動分類して管理するものであり、結果として図７に示したグループテーブル（ＧＴＢＬ）４２３が自動生成される。

【００６１】ステップＳ３１では、本実施形態における文書データベースに格納されている意味ベクトルによって自己組織化マップ学習を行う。この自己組織化マップ学習処理は、多次元ベクトル表現の集合を２次元マップに変換して分類する手法として一般的に行われており、公知であるが、ここではその概要を図１３を用いて簡単に説明する。

【００６２】図１３は自己組織化マップの学習の例を示す図である。図１３において、１はユニットＵ、２はＵの近傍ユニットＵｃである。図１３に示したように、自己組織化マップは正方形であり、Ｎ×Ｎ個（Ｎは正の整数）のユニットで構成される。各ユニットは、意味ベクトルと同次元のパターンベクトルを持つ。

【００６３】自己組織化マップの学習は、以下の手順で行う。まず、マップ中のユニットが持つ全てのパターンベクトルをゼロベクトルに初期化する。次に、文書データベース中の各意味ベクトルに対して、なす角が最小であるパターンベクトルを持つユニットＵを決定する。次に、Ｕとその近傍ユニットＵｃが持つパターンベクトルを意味ベクトルに近づける。このとき、各ユニットが持つパターンベクトルに対して、ユニットＵとの距離が近いほど、パターンベクトルを意味ベクトルに近づける度合いを高くする。初期化以外の以上の操作を予め決められた回数繰り返すと、２次元マップ上の各ユニットが持つベクトルパターンはある定値に収束し、自己組織化マップ学習が終了する。自己組織化マップ学習処理を終えると、ステップＳ３２へ進む。

【００６４】ステップＳ３２では、ステップＳ３１で学習が済んだ自己組織化マップ上のユニットに文書データベースに登録されている各文書の意味ベクトルを配置する処理が行われる。各文書の意味ベクトルは、マップ学習時と同様に、パターンベクトルとのなす角が最小であるユニットに配置される。処理を終えると、ステップＳ３３へ進む。

【００６５】ステップＳ３３では、ステップＳ３２によって２次元マップ上に配置された各文書データを８連結手法によってグループ化し、グループを決定する処理が行われる。ここで、８連結手法とは、２次元マップ上で縦横斜めに隣接するユニット同士を同じグループとしてまとめる手法である。図１４は、２次元マップ上において８連結手法によってグループを決定する例を示す図である。図１４において、３はグループＩＤ＝１のグループ、４はグループＩＤ＝２のグループ、５はグループＩＤ＝３のグループ、６はグループＩＤ＝４のグループである。なお、マップ上の数字は配置された文書データの数を示す。処理を終えると、ステップＳ３４へ進む。ステップＳ３４では、ステップＳ４３で決定したグループに属する文書データの全ての意味ベクトルとのなす角度が最小となる多次元ベクトルを求め、これをグループの意味ベクトルとする。

【００６６】ステップＳ３５では、ステップＳ３４で決定された意味近似文書グループのグループ名を決定する、グループ名決定処理が行われる。この処理は、図１５を用いて後述する。グループ名決定処理を終えるとステップＳ３６へ進む。ステップＳ３６は、ステップＳ３３～ステップＳ３５で作成されたグループの情報を参照して、グループテーブル（ＧＴＢＬ）４２３を作成する処理である。処理を終えると、意味近似文書グループ作成処理を終了する。なお、本例では、グループテーブル（ＧＴＢＬ）４２３において、文書データの総数の多い順に、グループＩＤを１から順に割り当てる。例えば、図１４において、グループ３のマッチング数は１５＋２８＋９＝５２で最も多く、図７に示すように、ＩＤ＝１が割り当てられている。なお、グループ作成処理の際には、図６の検索結果は破棄されているものとする。

【００６７】図１５は、ステップＳ３４におけるグループ名決定処理の詳細を説明するフローチャートである。図１５において、ステップＳ４１では、グループ名を決定するグループの意味ベクトル７０５と、意味概念辞書（ＭＤＩＣ）中の意味ベクトルとのマッチングを取り、ある閾値以上の類似度を持つ単語を当該グループのグループ名候補として複数個検索する。また、ここで用いる閾値は、予め本実施形態のシステムにおいて定まっているものとする。

【００６８】次に、ステップＳ４２において、ステップＳ４１で選択したグループ名候補単語が、シソーラスツリー上においてどの位置にあるかを判定し、シソーラスツリー上の中間部に位置する単語をグループ名として選

択する。これは、シソーラスツリー上の上位（根の部分）に位置する単語をグループ名として選択すると、意味概念が大きくなりすぎるためそのグループの持つ概念がぼやけてしまうのでそれを避けるためである。また、逆に、シソーラスツリー上の下位（子葉）部分に位置する単語をグループ名として選択すると、その単語が持つ意味概念が狭すぎるために、グループ名にふさわしくない文書データがグループに紛れ込む可能性が高くなるので、子葉部分に位置する単語は候補から外すためである。

【００６９】この処理において、シソーラスツリー上の中間部に位置する単語のうち、グループの意味ベクトル７０５との類似度が最も高い意味ベクトルを持つ単語がグループ名として決定される。処理を終えると、グループ名決定処理を終了する。

【００７０】以上のように、本実施形態によれば、文書データを登録する際に、文書の意味内容を示す多次元ベクトルが作成され、このベクトルを利用して予め文書が意味概念によって自動分類され、意味概念の類似している文書グループを表現するのにふさわしい単語がグループ名として付加される。そして、検索時のユーザクエリにこの単語が現れた場合に、検索結果候補をこの単語をグループ名とする意味近似グループに絞り込むことによって、意味ベクトルのマッチング処理量を減らし、迅速な検索を実行することが可能となる。また、特に、グループ名としてシソーラスツリーの中間に位置する単語を選択することにより、検索精度を高めることが可能となる。

【００７１】なお、本発明は上述した実施形態に限定されるものではない。例えば、上述の実施形態では、文書検索装置のバスとしてＰＣＩバスを採用しているが、ＩＳＡバスやＶＬバスなどでもまったく同様な文書検索装置を構成することが可能である。また、上述の実施形態では、ＯＳはＤＩＳＫに格納されているが、ＯＳをＲＯＭ上に格納しても同様な処理を行うことが可能である。

【００７２】また、上述の実施形態では、ＤＶＤ－ＭＥＤＩＡから文書検索処理プログラムおよび関連データを直接ＲＡＭにロードして実行させる例を示したが、このほかにＤＣＤ－ＭＥＤＩＡから文書から文書検索処理プログラムおよび関連データを一旦ＤＩＳＫに格納（インストール）しておき、本文書検索処理プログラムを動作させるときにＤＩＳＫからＲＡＭにロードするようにすることも可能である。

【００７３】また、上述の実施形態では、本文書検索処理プログラムを記憶する媒体としてＤＶＤ－ＭＥＤＩＡを用いているが、それ以外にＣＤ－ＭＥＤＩＡ，ＭＯ，ＤＦ，ＩＣメモリカード、光磁気カードなどを用いても良い。更に文書検索処理プログラムをＲＯＭに記憶しておき、これをメモリマップの一部となすように構成し、直接ＣＰＵで実行することも可能である。

【００７４】また、本発明の目的は、前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体（または記録媒体）を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は本発明を構成することになる。また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているオペレーティングシステム（ＯＳ）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００７５】さらに、記憶媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張カードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張カードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００７６】その他、本発明はその要旨を逸脱しない範囲で種々変形して実施することができる。

【００７７】

【発明の効果】以上説明したように、本発明によれば、多次元ベクトルマッチングを用いた文書検索処理において、膨大な文書データが検索対象となった場合でもその処理時間を短く保つことができる。

【図面の簡単な説明】

【図１】本実施形態による文書検索処理を実行するコンピュータ装置１００の構成を示すブロック図である。

【図２】本文書検索処理プログラム及び関連データをＤＶＤ－ＭＥＤＩＡ１０６に記録したときのＤＶＤ－ＭＥＤＩＡ１０６の内容の構成図である。

【図３】コンピュータ装置１００と本文書検索処理プログラムが記録されたＤＶＤ－ＭＥＤＩＡ１０６の模式図である。

【図４】本文書検索処理プログラムがＲＡＭ１０４にロードされ実行可能となった状態のメモリマップを示す図である。

【図５】ＣＯＵＮＴ＝１０の場合の検索結果バッファ（ＲＢＵＦ）４２２の構成例を説明した図である。

【図６】本実施形態における検索結果バッファ（ＲＢＵＦ）の構成例を説明した図である。

【図７】本実施形態におけるグループテーブル（ＧＴＢＬ）の構成例を説明した図である。

【図８】本実施形態における文書データベースに格納されている一つのデータ構造の構成例を説明した図である。

【図９】本実施形態における文書検索処理の全体を説明するフローチャートである。

【図１０】図９のステップＳ３における文書登録処理を詳細化したフローチャートである。

【図１１】図９に示したステップＳ４の文書検索処理の詳細を示すフローチャートである。

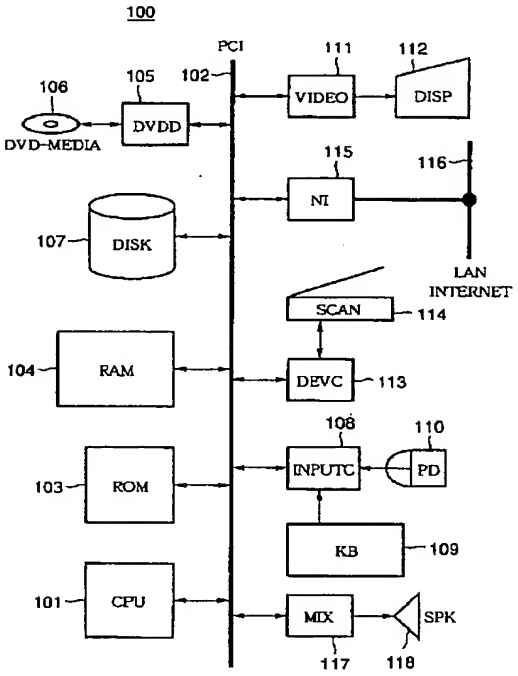
【図１２】図９のステップＳ６に示した意味近似文書グループ作成処理の詳細を示すフローチャートである。

【図１３】自己組織化マップの学習の例を示す図である。

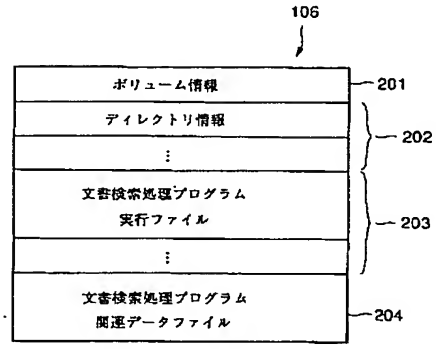
【図１４】自己組織化マップの学習において、２次元マップ上で８連結手法によってグループを決定する例を示す図である。

【図１５】ステップＳ３４におけるグループ名決定処理の詳細を説明するフローチャートである。

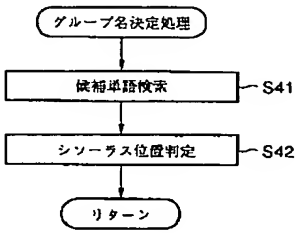
【図１】



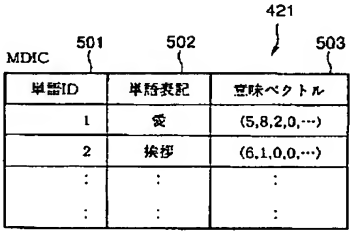
【図２】



【図１５】



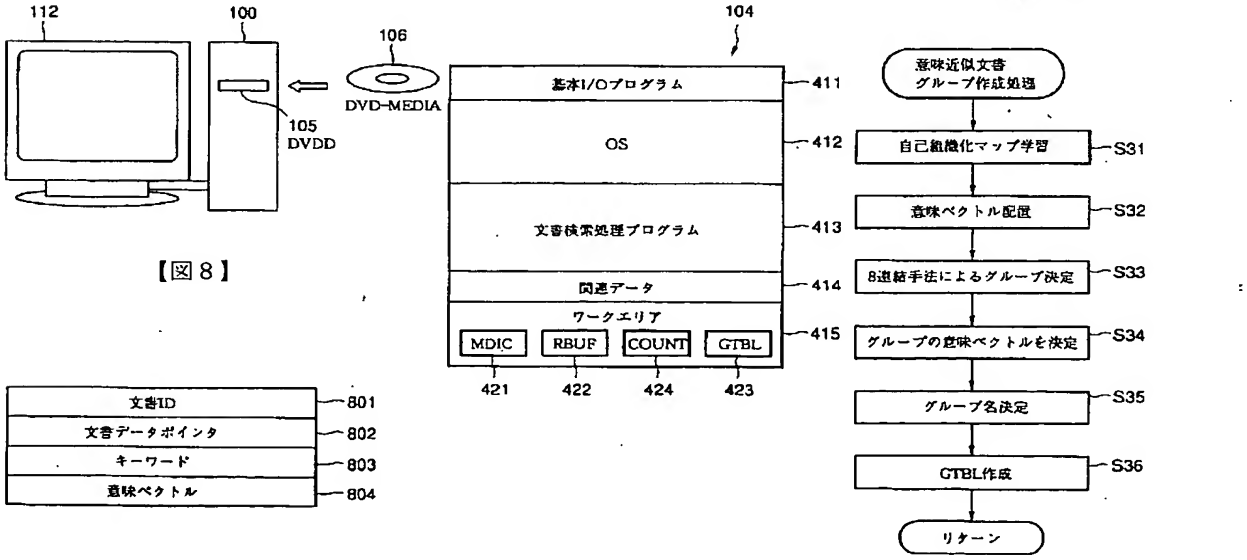
【図５】



【図３】

【図４】

【図１２】



【図８】

【図６】

422

RBUF

601 順位	602 意味近似度	603 意味ベクトル	604 文書ID	605 グループID
1	80	(1,0,4,8,…)	21365	1
2	75	(2,0,0,3,…)	90235	1
:	:	:	:	:
:	:	:	:	:
10	35	(0,0,0,5,…)	38106	3

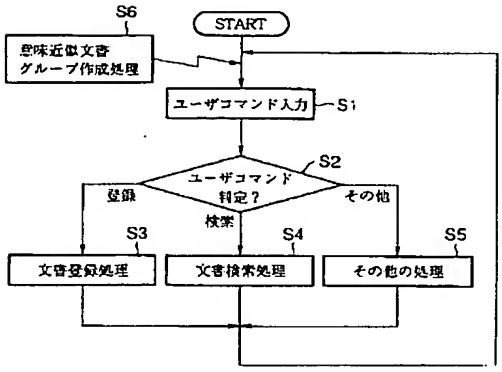
【図７】

423

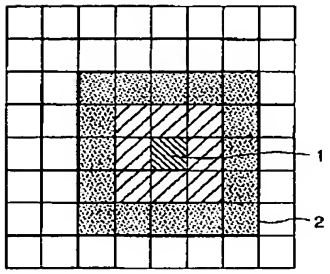
GTBL

701 グループID	702 グループ名	703 データ数	704 文書IDリスト	705 意味ベクトル
1	インターネット	52	21365, 90235, …	
2	モバイル	39	28215, 35671, …	
3	液晶	25	59321, 38106, …	
4	プリンタ	14	67391, 984, …	
:	:	:	:	
:	:	:	:	

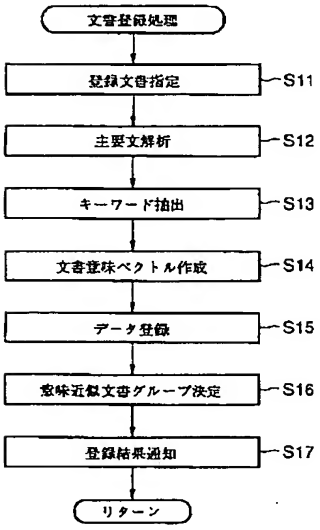
【図９】



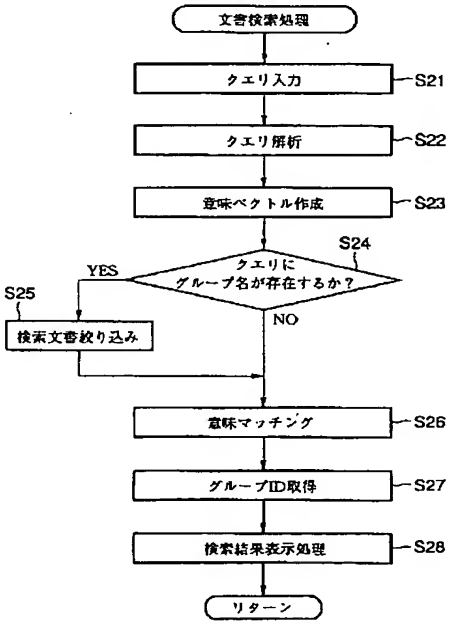
【図１３】



【図１０】



【図１１】



【図１４】